An information-theoretic framework for resolving community structure in complex networks

Martin Rosvall* and Carl T. Bergstrom†

Department of Biology, University of Washington, Seattle, WA 98195-1800

This paper develops a rigorous foundation for the concept of modularity in networks. We present an explicit method to best partition a complex network by finding an optimal compression of its topology, capitalizing on regularities in its structure. To illustrate the power of our approach, we partition a number of real-world and model networks and explain why our results outperform other methods.

We can comprehend the structure of a dauntingly complex network by identifying the communities or modules of which it is composed (1, 2, 3, 4, 5). When we describe a network as a set of interconnected modules, we are highlighting certain regularities of the network's structure while filtering out the relatively unimportant details. Thus a modular description of a network is a lossy compression of that network's topology — and the problem of community identification is at its core a problem of finding an efficient compression of the network's structure.

This view suggests that we should approach the challenge of identifying the community structure of a complex network as a fundamental problem in information theory (6, 7, 8). We provide the groundwork for this information-theoretic approach to community detection, and explain why this approach outperforms other methods for community detection.

Figure 1 illustrates our basic framework for identifying communities. We envision the process of describing a complex network by a simplified summary of its module structure as a communication process. The link structure of a complex network is a random variable X; a signaler knows the full form of the network X, and aims to convey much of this information in a reduced fashion to a signal receiver. To do so, the signaler encodes information about X as some simplified description Y. She sends the encoded message through a noiseless communication channel. The signal receiver observes the message Y, and then "decodes" this message, using it to make guesses Z about the structure of the original network X.

There are many different ways to describe a network X by a simpler description Y. Which of these is best? The answer to this question of course depends on what you want to do with the description. Nonetheless, information theory offers an appealing general answer to this question. Given some set of candidate descriptions Y_i, the best description Y of a random variable X is the one that tells the most about X — that is, the one that maximizes the mutual information $I(X;Y)$ between description and network.

Since we are interested in identifying community structure, we will explore descriptions Y that summarize the structure of a network X by enumerating the communities or modules within X, and describing the relations among them. In this paper, we will consider one particular method of encoding the community structure of X. More generally one could and indeed should consider alternative "encoders" so as to choose one best suited for the problem at hand.

We consider an unweighted and undirected network X of size n with l links, which can be described by the adjacency matrix

$$\mathbf{A}_{ij} = \begin{cases} 1 & \text{if there is a link between nodes } i \text{ and } j \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

We choose the description

$$Y = \left\{ \mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}, \mathbf{M} = \begin{pmatrix} l_{11} & \cdots & l_{1m} \\ \vdots & \ddots & \vdots \\ l_{m1} & \cdots & l_{mm} \end{pmatrix} \right\} \quad (2)$$

for m modules, where $\mathbf{a}$ is the module assignment vector, $a_i \in \{1, 2, \ldots, m\}$, and $\mathbf{M}$ is the module matrix. Given the assignment vector and the actual network, $\mathbf{M} = \mathbf{M}(X, \mathbf{a})$ is the modular description of the actual network, with the m modules connected pairwise by l_{ij} links between module i and j, and n_i nodes in module i connected internally by l_{ii} (see Fig. 1).

To find the best assignment $\mathbf{a}^*$ we now maximize the mutual information over all possible assignments of the nodes into m modules

$$\mathbf{a}^* = \arg\max_{\mathbf{a}} I(Y;X). \quad (3)$$

By definition, the mutual information $I(Y;X) = H(X) - H(X|Y) = H(X) - H(Z)$, where $H(X)$ is the information necessary to describe X and the conditional information $H(X|Y) = H(Z)$ is the information necessary to describe X given Y (see Fig. 1). We therefore seek to to minimize $H(Z)$. This is equivalent to constructing an assignment vector such that the set of network estimates Z in Fig. 1 is as small as possible. Given that the description Y assigns nodes to m modules,

$$H(Z) = \log\left[\prod_{i=1}^{m} \binom{n_i(n_i-1)/2}{l_i} \prod_{i>j} \binom{n_i n_j}{l_{ij}}\right], \quad (4)$$

where the parentheses denote the binomial coefficients and the logarithm is taken in base 2. Each of the m binomial

*Electronic address: rosvall@u.washington.edu; URL: http://www.tp.umu.se/~rosvall/
†URL: http://octavia.zoology.washington.edu/

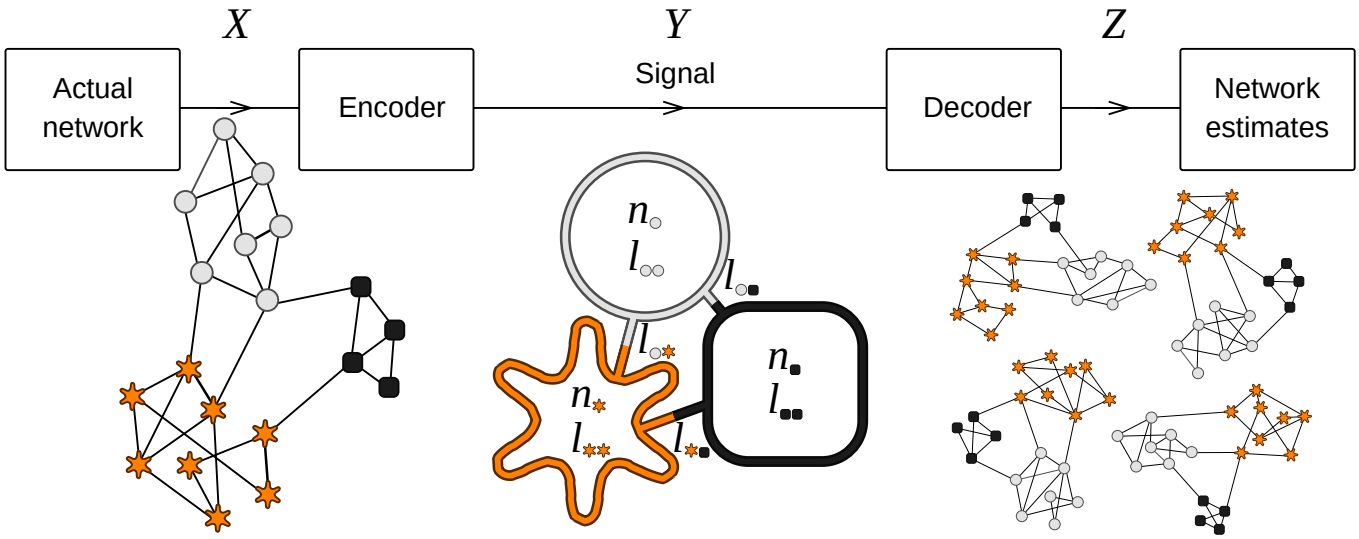

FIG. 1 Basic framework for detecting communities as a communication process. A signaler knows the full network structure and wants to send as much information as possible about the network to a receiver over a channel with limited capacity. The signaler therefore encodes the network into modules in a way that maximizes the amount of information about the original network. This figure illustrates an encoder that compresses the network into 3 modules $i = \circ, \blacksquare, \ast$, with n_i nodes and l_{ii} links in each module, and l_{ij} links between the modules. The receiver can then decode the message and construct a set of possible candidates for the original network. The smaller the set of candidates, the more information the signaler has managed to transfer.

coefficients in the first product gives the number of different modules that can be constructed with n_i nodes and l_{ii} links. Each of the $m(m+1)/2$ binomial coefficients in the second product gives the number of different ways module i and j can be connected to one another.

Figure 2 shows that our cluster-based compression method splits the network close to the division along which the actual dolphin groups were observed to split (9). Only 6 links cross between the two clusters, one with 21 members and the other with 41. Because it is computationally infeasible to check all possible partitions of even modestly-sized networks, we use Markov chain Monte Carlo (MCMC) with Metropolis-Hastings sampling to search for the partition that maximizes the mutual information between the description and the original network. We have confirmed the results with exhaustive search in the vicinity of the MCMC solutions.

We compare our results with the partition obtained by using the *modularity* approach introduced by Newman and Girvan in ref. (10); that technique has been widely adopted because of its appealing simplicity, its strong performance in benchmark tests (5), and the availability of powerful numerical techniques for dealing with large networks (11, 12, 13). Given a partitioning into m modules, the modularity Q is the sum of the contributions from each module i

$$Q = \sum_{i=1}^{m} l_{ii}/l - (d_i/2l)^2, \qquad (5)$$

where l_{ii} is the number of links between nodes in the i-th module, d_i the total degree in module i, and l is the total number of links in the network. When we maximize the modularity, we are not just minimizing the number of links between modules, but rather finding a configuration which maximizes the number of links within modules in the actual network minus the expected number of links within comparable modules in a random network with the same degree sequence. Or equivalently, we aim to divide the network such that the number of links within modules is higher than expected (12).

This approach works perfectly for networks where the modules are similar in size and degree sequence (5). However, when we partition the dolphin network in Fig. 2 using the modularity approach, we divide the network such that 12 instead of 6 links connect the two modules. Why? Because of the $(2l)^2$ denominator in the second term of Eq. (5), the choice of partition is highly sensitive to the total number of links in the system. Thus modularity-based partitioning is somewhat arbitrary in the sense that if we add an additional distinct cluster and then partition the whole network allowing one more module, the division point between the original modules may shift substantially with the increased number of links in the total system. For example, imagine that in the dolphin network in Fig. 2, we added a distinct group of dolphins forming a third cluster. If the third group is equal in size as the original network, the resulting three-way partition would divide the original part of the dolphin network just as our cluster-based compression method does.

To test our cluster-based compression method quantitatively we conducted the benchmark tests described in refs. (1, 5). In these tests, 128 nodes are divided into four equally sized groups with average degree 16. As the average number of links k_{out} from a node to nodes in other groups increases, it becomes harder and harder to identify the underlying group structure. We generated 100 different networks with the described procedure for each value $k_{\text{out}} = 6, 7, 8$ and searched with the method described above for the solution that max-

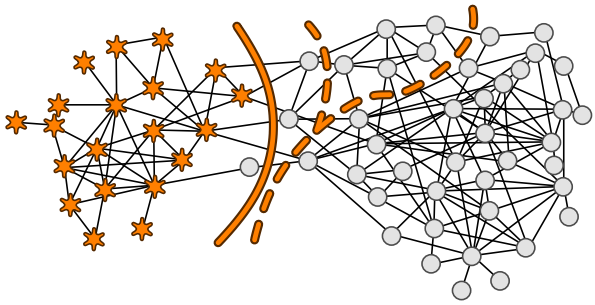

FIG. 2 The dolphin network by (9) partitioned with our cluster-based compression (solid line) and based on the modularity (dashed line). The stars and circles represent the two observed groups of dolphins. The right branch of the dashed line represents a split based on maximizing the modularity, which is different from the left branch solution based on the spectral analysis presented in ref. (14). The edge-betweenness algorithm presented in ref. (1) splits the network in the same way as our cluster-based compression method (10).

imized the mutual information between the description and the original network. We found a correct assignment with probability 1.00 (0) for $k_{\text{out}} = 6$, 0.98 (.01) for $k_{\text{out}} = 7$, and 0.89 (.05) for $k_{\text{out}} = 8$, with the standard deviation in the parenthesis. This accuracy is at the same level as a simulated annealing approach to maximize the modularity (2, 15) and outperforms all other approaches (5). The close accord between the results for the information-theoretic approach presented here and the best algorithm based on the modularity show that the two methodologies behave similarly when the module structure of the network is sufficiently regular.

In some special cases we will know a priori how many modules compose our sample network, but in general the task of resolving community structure is twofold. We must determine the number of modules in the network, and then we need to partition the nodes into that number of modules. The catch is that we cannot determine the optimal number of modules without also considering the assignments of nodes — so these problems need to be solved simultaneously. Below, we provide a solution grounded in algorithmic information theory.

Looking back at Fig. 1, the encoder seeks to find an compression of the network so that the decoder can make the best possible estimate of the actual network. One approach would be to have the encoder partition the network into n modules, one for each node, and thereby ensure that decoder can reconstruct the network completely, but under this approach nothing is gained either in compression or module identification. Therefore the encoder must balance the amount of information necessary to describe the network in modular form, as given by the signal Y in Fig. 1, and the uncertainty that remains once the decoder receives the modular description, as given by the size of the set of network estimates Z in Fig. 1. This is an optimal coding problem and can be resolved by the Minimum Description Length (MDL) principle (6, 17, 18). The idea is to exploit the regularities in the structure of the actual network X to summarize it in condensed form, without overfitting it. What do we mean by overfitting in this context?

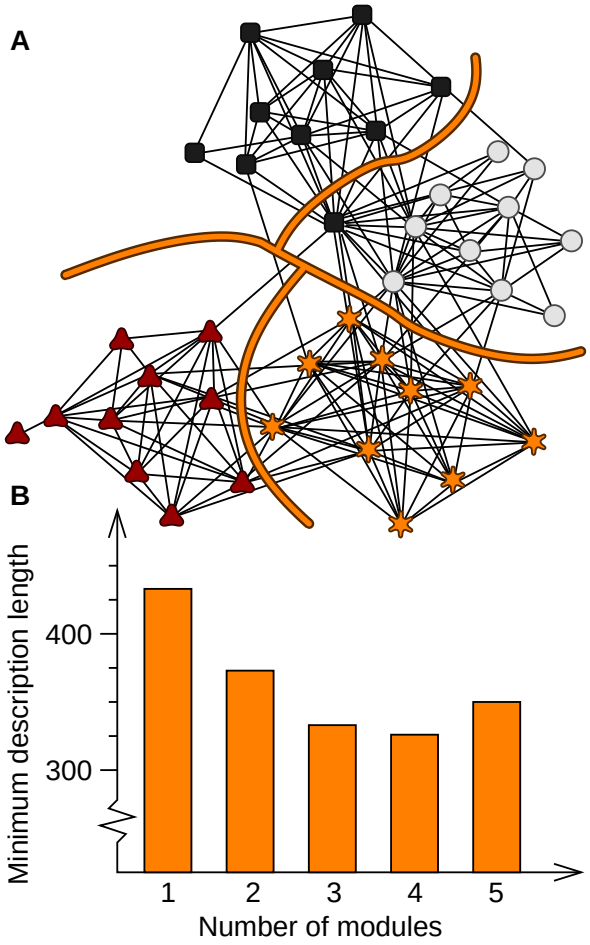

FIG. 3 Partitioning into an optimal number of modules. The network in panel A consists of 40 journals as nodes from four different fields: multidisciplinary physics (squares), chemistry (circles), biology (stars), and ecology (triangles). The 189 links connect nodes if at least one article from one of the journals cites an article in the other journal during 2004 (16). We have selected the 10 journals with the highest impact factor in the four different fields, but disregarded journals classified in one or more of the other fields. Panel B shows the minimum description length for the network in panel A partitioned into 1 to 5 different modules. The optimal partitioning into four modules is illustrated by the lines in panel A.

Figure 3 illustrates. We want to choose a set of modules for the journal citation network in Fig. 3 such that if we were to repeat the experiment next year, each journal would likely be assigned to the same module again. If we overfit the data, we may capture more of a specific year's data, but unwittingly also the noise that will not recur in next year's data.

To minimize the description length of the original network X, we look for the number of modules m such that the description length of the modular description Y plus the conditional description length — the amount of additional information that would be needed to specify X exactly to a receiver who had already decoded the description Y — is as short as possible (6). That is, we seek to minimize the sum

$$L(Y) + L(X|Y), \tag{6}$$

where $L(Y)$ is the length in bits of the signal and $L(X|Y)$ is number of bits needed to specify which of the network estimates implied by the signal Y is actually realized. The description length is easy to calculate in this discrete case and is given by

$$L(Y) + L(Z|Y) = n \log m + \frac{1}{2} m(m+1) \log l + H(Z), \qquad (7)$$

where the first and second term give the size necessary to encode the assignment vector $\mathbf{a}$ and the module matrix $\mathbf{M}(X, \mathbf{a})$, and $H(Z)$ is given in Eq. (4). Figure 3B shows the description length with the journal network partitioned into one to five modules. Four modules yield the minimum description length and we show the corresponding partition in Fig. 3A.

This cluster-based compression assigns 39 of the 40 journals into the proper categories, but places the central hub *Physical Review Letters* (PRL) in the chemistry cluster. This may seem like a mistake, given that PRL has 9 links to physics and only 8 to chemistry. Indeed, a partitioning based on the modularity score Q places PRL among the physics journals. But whatever its subject matter, the structural role that PRL plays in the journal network is really that of a chemistry journal. Like most of the chemistry journals, and unlike its compatriots in physics, PRL is closely linked to biology and somewhat connected to ecology.

While doing so yields a somewhat longer description length, we can also partition the network into two, three, or five modules. When we compress the network into two components, physics clusters together with chemistry and biology clusters together with ecology. When we split into three components, ecology and biology separate but physics and chemistry remain together in a single module. When we try to split the network into five modules, we get essentially the same partition as with four, only with the singly connected journal *Conservation Biology* split off by itself into its own partition. We cannot partition the network into more than five modules without creating at least one module that has a majority of its links to nodes in other modules. For purposes of illustration we have softened this constraint and accepted single nodes with only one out-link as modules. Because of this concept of what a module is (19), we constrained our model description to clusters satisfying $l_{ii} > l_{ij}$ for all i and j in Eq. (2). In many cases we get a higher mutual information if we remove this constraint; in such cases we typically observe that hubs are clustered together and peripheral nodes are clustered together. When this is true, we can describe the network structure more efficient by clustering nodes with similar roles instead of clustering nodes that are closely connected to one another. The mixture model approach provides an alternative method of identifying aspects of network structure beyond positive assortment (20).

To visualize the difference we split Zachary's classic karate club network (21) with (panel A) and without (panel B) the link constraint (Fig. 4). In A the partitioning corresponds exactly to the splitting that was observed by Zachary, but in B instead the 5 members with the highest degrees are clustered together. The compression with the hubs in one cluster and the peripheral nodes in the other cluster is in this case more efficient.

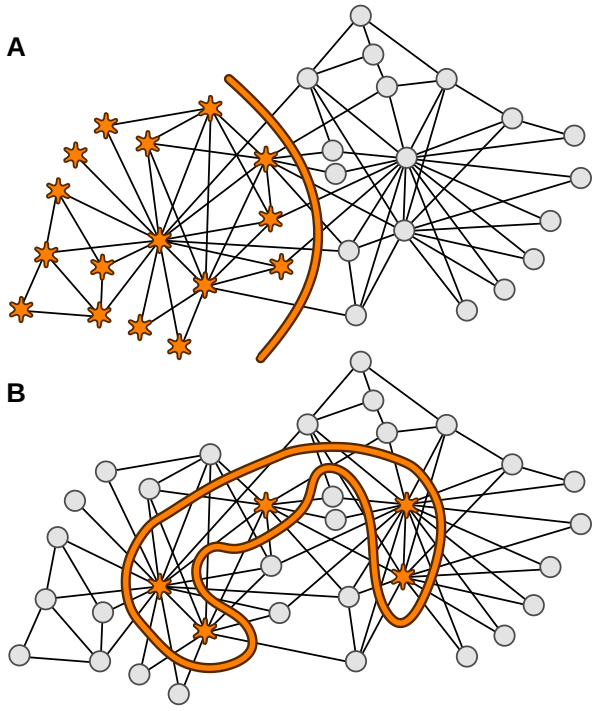

FIG. 4 Zachary's karate club network (21) partitioned into two modules based on the maximum mutual information with (panel A) and without (panel B) the link constraint. The partitioning with more links within modules than between modules in panel A clusters closely connected nodes together and the unconstrained partitioning in panel B clusters nodes with similar roles together.

Conclusions

We have shown that the process of resolving community structure in complex networks is at its core a problem in data compression. By drawing out the relationship between module detection and optimal coding we are able to ground the concept of network modularity in the rigorous formalism provided by information theory.

Enumerating the modules in a network is an act of description; there is an inevitable tradeoff between capturing most of the network structure at the expense of needing a long description with many modules, and omitting some aspects of network structure so as to allow a shorter description with fewer modules. Moreover, our information-theoretic approach suggests that there is a natural scale on which to best describe the network, a single answer to how to best balance this tradeoff between under- and over-description.

While the information-theoretic view described here provides a fully general basis for how to get the most information out of a network structure, the sort of information that we wish to extract may vary from application to application. In this paper, we have presented one particular encoder, which extracts information about the community structure of unweighted, undirected networks. However, the method is easy to general-

ize to directed and weighted networks as well, by modifying the modular description appropriately. To illustrate the power of our cluster based-compression, we have partitioned a number of real-world and model networks and explained why our results outperform other methods.

Acknowledgments

We thank Ben Althouse for generating the network used in Fig. 3. This work was supported by the National Institute of General Medical Sciences Models of Infectious Disease Agent Study program cooperative agreement 5U01GM07649.

References

1. M. Girvan, M. E. J. Newman, *Proc Natl Acad Sci USA* **99**, 7821 (2002).
2. R. Guimerà, L. A. N. Amaral, *Nature* **433**, 895 (2005).
3. P. Holme, M. Huss, H. Jeong, *Bioinformatics* pp. 532–538 (2003).
4. M. E. J. Newman, *Eur Phys J B* **38**, 321 (2004).
5. L. Danon, A. Díaz-Guilera, J. Duch, A. Arenas, *J Stat Mech* p. P09008 (2005).
6. J. Rissanen, *Automatica* **14**, 465 (1978).
7. C. E. Shannon, W. Weaver, *The mathematical theory of communication* (Univ of Illinois Press, 1949).
8. E. Ziv, M. Middendorf, C. H. Wiggins, *Phys Rev E* **71**, 046117 (2005).
9. D. Lusseau, *et al.*, *Behav Ecol Sociobiol* **54**, 396 (2003).
10. M. E. J. Newman, M. Girvan, *Phy Rev E* **69**, 026113 (2004).
11. M. E. J. Newman, *Phys Rev E* **69**, 066133 (2004).
12. M. E. J. Newman, *Proc Natl Acad Sci USA* **103**, 8577 (2006).
13. J. Reichardt, S. Bornholdt, *Phy Rev E* **74**, 016110 (2006).
14. M. E. J. Newman, *Phys Rev E* **74**, 036104 (2006).
15. R. Guimerà, L. A. N. Amaral, *J Stat Mech* p. P02001 (2005).
16. Journal citation reports (2004). Thompson Scientific, Institute for Scientific Information.
17. A. Barron, J. Rissanen, B. Yu, *IEEE T Inform Theory* **44**, 2743 (1998).
18. P. Grünwald, I. J. Myung, M. Pitt, eds., *Advances in minimum description length: theory and applications* (MIT Press, 2005), chap. 1–2.
19. F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, *Proc Natl Acad Sci USA* **101**, 2658 (2004).
20. M. E. J. Newman, E. A. Leicht, e-print physics/0611158.
21. W. W. Zachary, *J Anthropol Res* **33**, 452 (1977).